UNITED STATES PATENT AND TRADEMARK OFFICE

m-/

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/606,061	06/25/2003	Ross Cutler	302975.1	1637

7590 02/13/2007

Katrina A. Lyon
LYON & HARR, LLP
Suite 800
300 Esplanade Drive
Oxnard, CA 93036

EXAMINER	
JACKSON, JAKIEDA R	
ART UNIT	**PAPER NUMBER**
2626	

SHORTENED STATUTORY PERIOD OF RESPONSE	MAIL DATE	DELIVERY MODE
3 MONTHS	02/13/2007	PAPER

Please find below and/or attached an Office communication concerning this application or proceeding.

If NO period for reply is specified above, the maximum statutory period will apply and will expire 6 MONTHS from the mailing date of this communication.

Office Action Summary	Application No.	Applicant(s)
	10/606,061	CUTLER ET AL.
	Examiner	Art Unit
	Jakieda R. Jackson	2626

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☐ Responsive to communication(s) filed on _____.

2a)☐ This action is **FINAL**. 2b)☒ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) *1-32* is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☒ Claim(s) *1-7,9-18 and 20-32* is/are rejected.

7)☒ Claim(s) *8 and 19* is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)☐ The specification is objected to by the Examiner.

10)☒ The drawing(s) filed on *02 September 2003* is/are: a)☒ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☐ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☒ Notice of References Cited (PTO-892)

2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)

3) ☒ Information Disclosure Statement(s) (PTO/SB/08) Paper No(s)/Mail Date _____.

4) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date. _____ .

5) ☐ Notice of Informal Patent Application

6) ☐ Other: _____.

DETAILED ACTION

Claim Objections

1. The numbering of claims must be numbered consecutively beginning with the number next following the highest numbered claims.

- Claim 15 was numbered twice. The claim following claim 13 has been renumbered to claim 14 followed by claim 1.
- Claim 23 is missing. Therefore claims 24-33 are renumbered to claims 23-32.

2. Claims 11 is objected to because of the following informalities:

Claim 11 depends from itself and does not refer to another preceding claim. For purposes of examination, claim 11 depends from claim 1.

3. Claims 14 is objected to because of the following informalities:

- Claim 14 dependency for examination purposes have been changed to depend from claim 13 instead of 14.

Appropriate correction is required.

Claim Rejections - 35 USC § 103

4. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

> (a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

5. **Claims 1-2, 9-10, 12, 24 and 28-31** are rejected under 35 U.S.C. 103(a) as being unpatentable over Maali et al. (USPN 6,567,775), hereinafter referenced as Maali in view of Stork (USPN 5,586,215), hereinafter referenced as Stork.

Regarding **claim 1**, Maali discloses a computer-implemented process for detecting speech, comprising the process actions of:

inputting associated audio and video training data containing a person's face that is periodically speaking (audio-video; column 3, lines 51-60);

wherein said training comprises the following process actions:

computing audio features from said audio training data wherein said audio feature is the energy over an audio frame (column 3, lines 51-60 with column 6, lines 5-24);

computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed (lip movements; column 1, lines 47-54); and

correlating said audio features and video features to determine when a person is speaking (column 3, lines 51-60), but does not specifically teach using said audio and video signals to train a time delay neural network to determine when a person is speaking.

Stork teaches an acoustic and visual speech recognition system using said audio and video signals to train a time delay neural network to determine when a person is speaking (column 4, lines 46-64 with column 7, lines 22-29), in order to accommodate

utterances that may be of variable length, as well as somewhat unpredictable in the time of utterance onset.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Maali's process wherein it uses said audio and video signals to train a time delay neural network to determine when a person is speaking, as taught by Stork, to obtain the most probable utterance (column 4, lines 47-61).

Regarding **claim 2**, Maali discloses a computer-implemented process for detecting speech in an audio-visual sequence, but does not specifically teach a process further comprising the process action of preprocessing the audio and video signals prior to using said audio and video signals to train a Time Delay Neural Network.

Stork teaches an acoustic and visual speech recognition system wherein a process further comprising the process action of preprocessing the audio and video signals prior to using said audio and video signals to train a Time Delay Neural Network (column 4, lines 46-64 with column 7, lines 22-29), in order to accommodate utterances that may be of variable length, as well as somewhat unpredictable in the time of utterance onset.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Maali's process teach wherein a process further comprising the process action of preprocessing the audio and video signals prior to using said audio and video signals to train a Time Delay Neural Network, as taught by Stork, to obtain the most probable utterance (column 4, lines 47-61).

Regarding **claim 9**, Maali discloses a process further comprising the process actions of:

inputting an associated audio and video sequence of a person periodically speaking (column 3, lines 51-60), but does not specifically teach using said trained Time Delay Neural Network to determine when in said audio and video sequence said person is speaking.

Stork teaches an acoustic and visual speech recognition system using said trained Time Delay Neural Network to determine when in said audio and video sequence said person is speaking (column 4, lines 46-64 with column 7, lines 22-29), in order to accommodate utterances that may be of variable length, as well as somewhat unpredictable in the time of utterance onset.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Maali's process wherein it uses said trained Time Delay Neural Network to determine when in said audio and video sequence said person is speaking, as taught by Stork, to obtain the most probable utterance (column 4, lines 47-61).

Regarding **claim 10**, Maali discloses a computer-implemented process for detecting speech in an audio-visual sequence, but does not specifically teach the process action of preprocessing the associated audio and video sequence prior to using said trained Time Delay Neural Network to determine if a person is speaking.

Stork teaches an acoustic and visual speech recognition system wherein the process action of preprocessing the associated audio and video sequence prior to using

said trained Time Delay Neural Network to determine if a person is speaking (column 4, lines 46-64 with column 7, lines 22-29), in order to accommodate utterances that may be of variable length, as well as somewhat unpredictable in the time of utterance onset.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Maali's process teach wherein a process further comprising the process action of preprocessing the associated audio and video sequence prior to using said trained Time Delay Neural Network to determine if a person is speaking, as taught by Stork, to obtain the most probable utterance (column 4, lines 47-61).

Regarding **claim 12**, Maali discloses a computer-readable memory containing a computer program that is executable by a computer to perform the process (column 4, lines 9-22).

Regarding **claim 24**, Maali discloses a computer-implemented process for detecting speech in an audio-visual sequence wherein more than one person is speaking at a time, comprising the process actions of:

inputting associated audio and video training data containing more than one person's face wherein each person is periodically speaking at the same time as the other person or persons (column 4, line 51 – column 4, lines 8); and

wherein said training comprises the following process actions:

computing audio features from said audio training data wherein said audio feature is the energy over an audio frame (column 3, lines 51-60 with column 6, lines 5-24);

computing video features from said video training signals to determine whether a given person's mouth is open or closed(lip movements; column 1, lines 47-54); and

correlating said audio features and video features to determine when a given person is speaking (column 3, lines 51-61).

Stork teaches an acoustic and visual speech recognition system using said audio and video signals to train a time delay neural network to determine which person is speaking at a given time (column 4, lines 46-64 with column 7, lines 22-29), in order to accommodate utterances that may be of variable length, as well as somewhat unpredictable in the time of utterance onset.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Maali's process wherein it uses said audio and video signals to train a time delay neural network to determine which person is speaking at a given time, as taught by Stork, to obtain the most probable utterance (column 4, lines 47-61).

Regarding **claim 28**, Maali discloses a computer-implemented process for detecting speech, comprising the process actions of:

inputting associated audio and video training data containing a person's face that is periodically speaking (column 3, lines 51-60); and

wherein said training comprises the following process actions:

computing audio features from said audio training (column 3, lines 51-60);

computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed (column 1, lines 47-54); and

correlating said audio features and video features to determine when a person is speaking (column 3, lines 51-60).

Stork teaches an acoustic and visual speech recognition system using said audio and video signals to train a statistical learning engine to determine when a person is speaking (column 4, lines 46-64 with column 7, lines 22-29), in order to accommodate utterances that may be of variable length, as well as somewhat unpredictable in the time of utterance onset.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Maali's process wherein it uses said audio and video signals to train a statistical learning engine to determine when a person is speaking, as taught by Stork, to obtain the most probable utterance (column 4, lines 47-61).

Regarding **claim 29**, Maali discloses a computer-implemented process wherein said audio feature is the acoustical energy over an audio frame (column 3, lines 51-60 with column 6, lines 5-24).

Regarding **claim 30**, Maali discloses a computer-implemented process for detecting speech, but does not specifically teach wherein said audio feature is defined by Mel cepstrum coefficients.

Stork discloses an acoustic and visual speech recognition system wherein said audio feature is defined by Mel cepstrum coefficients (column 6, lines 16-51), to generate a new sequence of numbers.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Maali's process wherein audio feature is defined by Mel cepstrum coefficients, as taught by Stork, to result in a different, but effective, dynamic visual data vector; column 5, lines 51-61).

Regarding **claim 31**, Maali discloses a computer-implemented process for detecting speech in an audio-visual sequence, but does not specifically teach wherein said statistical learning engine is a Time Delay Neural Network.

Stork teaches an acoustic and visual speech recognition system wherein said statistical learning engine is a Time Delay Neural Network (column 4, lines 46-64 with column 7, lines 22-29), in order to accommodate utterances that may be of variable length, as well as somewhat unpredictable in the time of utterance onset.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Maali's process teach wherein said statistical learning engine is a Time Delay Neural Network, as taught by Stork, to obtain the most probable utterance (column 4, lines 47-61).

6. **Claims 6-7** are rejected under 35 U.S.C. 103(a) as being unpatentable over Maali in view of Stork, as applied to claim 1, and in further view of Liang et al. (PGPUB 2003/0212552), hereinafter referenced as Liang.

Regarding **claim 6**, Maali discloses a process wherein the process action of computing video features from said video training signals comprises the process actions of:

using a face detector to locate a face in said video training signals (face detector; column 4, lines 1-8) and

using the geometry of a typical face to estimate the location of a mouth and extracting a mouth image (column 11, lines 59-66), but does not specifically teach stabilizing the mouth, using Linear Discriminant Analysis and designating values for the mouth.

Stork discloses an acoustic and visual speech recognition system comprising:

stabilizing the mouth image to remove any translational motion of the mouth caused by head movement (mouth rested; column 5, lines 37-61) and

designating values of mouth openness wherein the values range from -1 for the mouth being closed, to +1 for the mouth being open (assigned a value plus or minus; column 5, lines 37-61), to convey the essential visual information.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Maali's process wherein it stabilizes the mouth

and assigns values to the mouth, as taught by Stork, to result in a different, but effective, dynamic visual data vector (column 5, lines 51-61).

Maali in view Stork disclose a computer-implemented process for detecting speech, but does not specifically teach using a Linear Discriminant Analysis (LDA) projection to determine if the mouth in the segmented mouth image is open or closed.

Liang discloses audiovisual speech recognition using a Linear Discriminant Analysis (LDA) projection to determine if the mouth in the segmented mouth image is open or closed (column 2, paragraph 0015), to assign pixels in the mouth region to the lip and face classes.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Maali in view of Stork's process, wherein it uses a Linear Discriminant Analysis (LDA) projection to determine if the mouth in the segmented mouth image is open or closed, as taught by Liang, to find the best discrimination between the classes (column 2, paragraph 0015).

Regarding **claim 7**, Maali discloses a process for detecting speech, but does not teach a process wherein the process action of stabilizing the mouth image comprises the process action of using normalized cross correlation to remove any of said translational movement.

Stork discloses an acoustic and visual speech recognition system wherein the process action of stabilizing the mouth image (mouth rested position) comprises the process action of using normalized (normalization) cross correlation to remove any of

said translational movement (column 5, lines 37-62), to convey the essential visual information.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Maali's process wherein the process action of stabilizing the mouth image comprises the process action of using normalized cross correlation to remove any of said translational movement, as taught by Stork, to result in a different, but effective, dynamic visual data vector (column 5, lines 51-61).

7. **Claims 3-5 and 11,** are rejected under 35 U.S.C. 103(a) as being unpatentable over Maali, in view of Stork, as applied to claim 2 above, and in further view of Nefian et al. (PGPUB 2004/0122675), hereinafter referenced as Nefian.

Regarding **claim 3**, Maali in view of Stork disclose a process wherein said process action of preprocessing the audio and video signals comprises the process actions of:

segmenting the audio data signals (Stork; segment audio; column 4, lines 57-65);

segmenting the video data signals (Stork; segment video; column 4, lines 57-65);

extracting audio features (Stork; extract audio; column 6, lines 5-24); and

extracting video features (Stork; extract video; column 6, lines 5-24), but does not specifically teach reducing the noise of the audio signals.

Nefian discloses audiovisual continuous speech recognition system reducing the noise of the audio signals (column 3, paragraph 0026), to increase the recognition rate.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Maali in view of Stork's process, wherein it reduces the noise of the audio signals, as taught by Nefian, to reduce the parameter space and overall complexity (column 3, paragraph 0026).

Regarding **claim 4**, Maali discloses an audiovisual speech recognition process wherein the process action of segmenting the audio data signal comprises the process action of segmenting the audio data to determine regions of speech and non –speech (column 6, line 49 – column 7, line 54 with column 10, lines 58-65).

Regarding **claim 5**, Maali discloses a process wherein the process action of segmenting the video data signal comprises the process action of segmenting the video data to determine at least one face and a mouth region within said determined faces (column 11, line 59 – column 12, line 12).

Regarding **claim 11**, Maali in view of Stork disclose a process wherein said process action of preprocessing the audio and video sequence comprises the process actions of:

segmenting the audio data in said sequence (Stork; segment audio; column 4, lines 57-65);

segmenting the video data signals in said sequence (Stork; segment video; column 4, lines 57-65);

extracting audio features from said sequence (Stork; extract audio; column 6, lines 5-24); and

extracting video features from said sequence (Stork; extract video; column 6, lines 5-24), but does not specifically teach reducing the noise of the audio signals in said sequence.

Nefian discloses audiovisual continuous speech recognition system reducing the noise of the audio signals in said sequence (column 3, paragraph 0026), to increase the recognition rate.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Maali in view of Stork's process, wherein it reduces the noise of the audio signals in said sequence, as taught by Nefian, to reduce the parameter space and overall complexity (column 3, paragraph 0026).

8. **Claims 13-15, 20-22** are rejected under 35 U.S.C. 103(a) as being unpatentable over Bakis et al. (USPN 6,219,639), hereinafter referenced as Bakis in view of Stork.

Regarding **claim 13**, Bakis discloses a computer-readable medium having computer-executable instructions for use in detecting when a person in a synchronized audio video clip is speaking, said computer executable instructions comprising:

inputting one or more captured video and synchronized audio clips (synchronize lip movement with speech; column 2, lines 21-62),

segmenting (segment) said audio and video clips to remove portions of said video and synchronized (synchronize) audio clips not needed in determining if a

speaker in the captured video and synchronized audio clips is speaking (column 4, lines 11-67);

extracting audio and video features in said captured video and synchronized audio clips to be used in determining if a speaker in the captured (extracted attribute; abstract with column 4, lines 10-67); and wherein an audio feature is the energy over an audio frame and wherein said video feature is the openness of a person's mouth (column 10, lines 5-35), but does not specifically teach training a Time Delay Neural Network to determine when a person is speaking using said extracted audio and video features.

Stork teaches an acoustic and visual speech recognition system training a Time Delay Neural Network to determine when a person is speaking using said extracted audio and video features (column 4, lines 46-64 with column 7, lines 22-29), in order to accommodate utterances that may be of variable length, as well as somewhat unpredictable in the time of utterance onset.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Bakis' medium to train a Time Delay Neural Network to determine when a person is speaking using said extracted audio and video features, as taught by Stork, to obtain the most probable utterance (column 4, lines 47-61).

Regarding **claim 14**, Bakis discloses a medium for detecting speech, but does not specifically teach wherein the instruction for training a Time Delay Neural Network

further comprises a sub-instruction for correlating said audio features and video features to determine when a person is speaking.

Stork teaches an acoustic and visual speech recognition system wherein the instruction for training a Time Delay Neural Network further comprises a sub-instruction for correlating said audio features and video features to determine when a person is speaking (column 4, lines 46-64 with column 7, lines 22-29), in order to accommodate utterances that may be of variable length, as well as somewhat unpredictable in the time of utterance onset.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Bakis' medium wherein the instruction for training a Time Delay Neural Network further comprises a sub-instruction for correlating said audio features and video features to determine when a person is speaking, as taught by Stork, to obtain the most probable utterance (column 4, lines 47-61).

Regarding **claim 15**, Bakis discloses the computer-readable medium further comprising instructions for:

inputting a captured video and synchronized audio clip for which it is desired to detect a person speaking (column 4, lines 10-67), but does not specifically teach using said trained Time Delay Neural Network to determine when a person is speaking in the captured video and synchronized audio clip for which it is desired to detect a person speaking by using said extracted audio and video features.

Stork teaches an acoustic and visual speech recognition system using said trained Time Delay Neural Network to determine when a person is speaking in the

captured video and synchronized audio clip for which it is desired to detect a person speaking by using said extracted audio and video features (column 4, lines 46-64 with column 7, lines 22-29), in order to accommodate utterances that may be of variable length, as well as somewhat unpredictable in the time of utterance onset.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Bakis' medium wherein it uses said trained Time Delay Neural Network to determine when a person is speaking in the captured video and synchronized audio clip for which it is desired to detect a person speaking by using said extracted audio and video features, as taught by Stork, to obtain the most probable utterance (column 4, lines 47-61).

Regarding **claim 20**, Bakis discloses a system for detecting a speaker in a video segment that is synchronized with associated audio, the system comprising:

a general purpose computing device (column 10, lines 5-35); and

a computer program comprising program modules executable by the computing device, wherein the computing device is directed by the program modules of the computer program to (column 10, lines 5-35),

input one or more captured video and synchronized audio segments (column 2, lines 21-47 with column 4, lines 10-67),

segment said audio and video segments to remove portions of said video and synchronized audio segments not needed in determining if a speaker in the captured video and synchronized audio segments is speaking (column 4, lines 10-67);

extract audio and video features in said captured video and synchronized audio segments to be used in determining if a speaker in the captured video and synchronized audio segments is speaking, wherein said audio feature is the energy over an audio frame and said video feature is the openness of a person's mouth in said video and synchronized audio segments (column 4, lines 10-67); and

input a captured video and synchronized audio clip for which it is desired to detect a person speaking (column 4, lines 10-67), but does not specifically teach training a TDNN.

Stork teaches an acoustic and visual speech recognition system training a Time Delay Neural Network to determine when a person is speaking using said extracted audio and video features and use said trained Time Delay Neural Network to determine when a person is speaking in the captured video and synchronized audio segments for which it is desired to detect a person speaking (column 4, lines 46-64 with column 7, lines 22-29), in order to accommodate utterances that may be of variable length, as well as somewhat unpredictable in the time of utterance onset.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Bakis' medium wherein it trains a Time Delay Neural Network to determine when a person is speaking using said extracted audio and video features and use said trained Time Delay Neural Network to determine when a person is speaking in the captured video and synchronized audio segments for which it is desired to detect a person speaking, as taught by Stork, to obtain the most probable utterance (column 4, lines 47-61).

Regarding **claim 21**, Bakis discloses a system wherein it outputs a 1 when a person is talking for each frame in said captured video and synchronized audio segments for which it is desired to detect a person speaking, and outputs a 0 when no person is talking (column 12, lines 12-65), but does not specifically teach using Time Delay Neural Network to train.

Stork teaches an acoustic and visual speech recognition system using Time Delay Neural Network to train (column 4, lines 46-64 with column 7, lines 22-29), in order to accommodate utterances that may be of variable length, as well as somewhat unpredictable in the time of utterance onset.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Bakis' medium wherein it using Time Delay Neural Network to train, as taught by Stork, to obtain the most probable utterance (column 4, lines 47-61).

Regarding **claim 22**, Bakis discloses a system wherein said Time Delay Neural Network comprises:

one output, wherein said output is set to 0 when no person in the video and synchronized audio segment is speaking; and wherein said output is set to 1 when a person in the video and synchronized audio segment is speaking (column 12, lines 12-65), but does not specifically teach an input layer and two hidden layers

Stork discloses an acoustic and visual recognition system comprising an input layer (column 8, lines 24-32) and two hidden layers (column 15, lines 12-37), in order to accommodate utterances that may be of variable length.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Bakis' system wherein it comprises an input layer and two hidden layers, as taught by Stork, to enhance understanding (column 1, lines 36-51).

9. **Claim 16** is rejected under 35 U.S.C. 103(a) as being unpatentable over Bakis in view of Stork, as applied to claim 13 above, and in further view of Nefian.

Regarding **claim 16**, Bakis in view of Stork disclose the computer-readable medium for detecting speech, but does not specifically teach a medium further comprising an instruction for reducing noise in said audio video clips prior to said instruction for segmenting said audio and video clips.

Nefian discloses audiovisual continuous speech recognition system reducing noise in said audio video clips prior to said instruction for segmenting said audio and video clips (column 3, paragraph 0026), to increase the recognition rate.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Bakis and view of Stork's process, wherein it reduces noise in said audio video clips prior to said instruction for segmenting said audio and video clips, as taught by Nefian, to reduce the parameter space and overall complexity (column 3, paragraph 0026).

10. **Claims 17-18 and 23** are rejected under 35 U.S.C. 103(a) as being unpatentable over Bakis in view of Stork and in further view of Liang.

Regarding **claim 17**, Bakis discloses a process wherein the process action of computing video features from said video training signals comprises the process actions of:

using a face detector to locate a face in said video training signals (column 2, lines 21-47 with column 4, lines 10-67) and

using the geometry of a typical face to estimate the location of a mouth and extracting a mouth image (column 6, lines 7-19), but does not specifically teach stabilizing the mouth, using Linear Discriminant Analysis and designating values for the mouth.

Stork discloses an acoustic and visual speech recognition system comprising:

stabilizing the mouth image to remove any translational motion of the mouth caused by head movement (mouth rested; column 5, lines 37-61) and

designating values of mouth openness wherein the values range from -1 for the mouth being closed, to +1 for the mouth being open (assigned a value plus or minus; column 5, lines 37-61), to convey the essential visual information.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Bakis' process wherein it stabilizes the mouth and assigns values to the mouth, as taught by Stork, to result in a different, but effective, dynamic visual data vector (column 5, lines 51-61).

Bakis in view Stork disclose a computer-implemented process for detecting speech, but does not specifically teach using a Linear Discriminant Analysis (LDA) projection to determine if the mouth in the segmented mouth image is open or closed.

Liang discloses audiovisual speech recognition using a Linear Discriminant Analysis (LDA) projection to determine if the mouth in the segmented mouth image is open or closed (column 2, paragraph 0015), to assign pixels in the mouth region to the lip and face classes.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Bakis in view of Stork's process, wherein it uses a Linear Discriminant Analysis (LDA) projection to determine if the mouth in the segmented mouth image is open or closed, as taught by Liang, to find the best discrimination between the classes (column 2, paragraph 0015).

Regarding **claim 18**, Bakis discloses the computer-readable medium for detecting speech, but does not specifically teach wherein said sub-instruction for stabilizing the mouth image to remove any translational motion of the mouth caused by head movement employs normalized cross correlation.

Stork discloses an acoustic and visual speech recognition system wherein said sub-instruction for stabilizing the mouth image to remove any translational motion of the mouth caused by head movement employs normalized cross correlation (mouth rested; column 5, lines 37-61), to convey the essential visual information.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Bakis' process wherein said sub-instruction for

stabilizing the mouth image to remove any translational motion of the mouth caused by head movement employs normalized cross correlation, as taught by Stork, to result in a different, but effective, dynamic visual data vector (column 5, lines 51-61).

Regarding **claim 23**, Bakis discloses a system wherein the module for extracting audio and video features comprises sub-modules to extract the video features comprising:

using a face detector to locate a face in said video training signals (column 2, lines 21-47 with column 4, lines 10-47);

using the geometry of a typical face to estimate the location of a mouth and extracting a mouth image (column 6, lines 7-19), but does not specifically teach stabilizing the mouth and using Linear Discriminant Analysis.

Stork discloses an acoustic and visual speech recognition system comprising:

stabilizing the mouth image to remove any translational motion of the mouth caused by head movement (mouth rested; column 5, lines 37-61), to convey the essential visual information.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Bakis's system wherein it stabilizes the mouth, as taught by Stork, to result in a different, but effective, dynamic visual data vector (column 5, lines 51-61).

Bakis in view Stork disclose a computer-implemented process for detecting speech, but does not specifically teach using a Linear Discriminant Analysis (LDA) projection to determine if the mouth in the segmented mouth image is open or closed.

Liang discloses audiovisual speech recognition using a Linear Discriminant Analysis (LDA) projection to determine if the mouth in the segmented mouth image is open or closed (column 2, paragraph 0015), to assign pixels in the mouth region to the lip and face classes.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Bakis in view of Stork's process, wherein it uses a Linear Discriminant Analysis (LDA) projection to determine if the mouth in the segmented mouth image is open or closed, as taught by Liang, to find the best discrimination between the classes (column 2, paragraph 0015).

11. **Claims 25-27** is rejected under 35 U.S.C. 103(a) as being unpatentable over Maali in view of Stork, as applied to claim 24 above, and in further view of Liang and in further view of Applicant's Admitted Prior Art (PGPUB 2004/0267521).

Regarding **claim 25**, Maali discloses a process wherein the process action of computing video features from said video training signals comprises the process actions of:

using a face detector to locate a face in said video training signals (face detector; column 4, lines 1-8) and

using the geometry of a typical face to estimate the location of a mouth and extracting a mouth image (column 11, lines 59-66), but does not specifically teach

stabilizing the mouth, using Linear Discriminant Analysis and designating values for the mouth.

Stork discloses an acoustic and visual speech recognition system comprising:

stabilizing the mouth image to remove any translational motion of the mouth caused by head movement (mouth rested; column 5, lines 37-61) and

designating values of mouth openness wherein the values range from -1 for the mouth being closed, to +1 for the mouth being open (assigned a value plus or minus; column 5, lines 37-61), to convey the essential visual information.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Maali's process wherein it stabilizes the mouth and assigns values to the mouth, as taught by Stork, to result in a different, but effective, dynamic visual data vector (column 5, lines 51-61).

Maali in view Stork disclose a computer-implemented process for detecting speech, but does not specifically teach using a Linear Discriminant Analysis (LDA) projection to determine if the mouth in the segmented mouth image is open or closed.

Liang discloses audiovisual speech recognition using a Linear Discriminant Analysis (LDA) projection to determine if the mouth in the segmented mouth image is open or closed (column 2, paragraph 0015), to assign pixels in the mouth region to the lip and face classes.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Maali in view of Stork's process, wherein it uses a Linear Discriminant Analysis (LDA) projection to determine if the mouth in the

segmented mouth image is open or closed, as taught by Liang, to find the best discrimination between the classes (column 2, paragraph 0015).

Maali in view of Stork and Liang disclose a process for detecting speech, but does not specifically teach using a microphone array beam form on each face.

However, based on Applicant's own admission beamforming is a well known technique for improving the sound quality of the speaker (columns 4-5, paragraph 0055).

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Maali in view of Stork and Liang's process wherein it teaches beamforming, as taught by Applicant, to improve the sound quality oif the speaker by filtering out sound not coming from the direction of the speaker (columns 4-5, paragraph 0055).

Regarding **claim 26**, Maali in view of Stork and Liang disclose a process for detecting speech, but does not specifically teach wherein said audio feature is computed using said beam formed audio training data.

However, based on Applicant's own admission beamforming is a well known technique for improving the sound quality of the speaker (columns 4-5, paragraph 0055).

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Maali in view of Stork and Liang's process wherein it teaches beamforming, as taught by Applicant, to improve the sound quality oif

the speaker by filtering out sound not coming from the direction of the speaker (columns 4-5, paragraph 0055).

Regarding **claim 27**, Maali discloses a process further comprising the process actions of:

inputting an associated audio and video sequence of a person periodically speaking (column 3, lines 51-60), but does not specifically teach using said trained Time Delay Neural Network to determine when in said audio and video sequence said person is speaking.

Stork teaches an acoustic and visual speech recognition system using said trained Time Delay Neural Network to determine when in said audio and video sequence said person is speaking (column 4, lines 46-64 with column 7, lines 22-29), in order to accommodate utterances that may be of variable length, as well as somewhat unpredictable in the time of utterance onset.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Maali's process wherein it uses said trained Time Delay Neural Network to determine when in said audio and video sequence said person is speaking, as taught by Stork, to obtain the most probable utterance (column 4, lines 47-61).

12. **Claim 32** is rejected under 35 U.S.C. 103(a) as being unpatentable over Maali in view of Stork, as applied to claim 28 above, and in further view of Nefian.

Regarding **claim 32**, Maali in view of Stork disclose a computer-implemented process wherein it detects speech, but does not specifically teach wherein said statistical learning engine is a Support Vector Machine.

Nefian discloses audiovisual continuous speech recognition wherein said statistical learning engine is a Support Vector Machine (column 2, paragraph 0016), to remove false alarms.

Therefore, it would have been obvious to one of ordinary skill in the art at the time the invention was made to modify Maali in view of Stork's process, wherein said statistical learning engine is a Support Vector Machine, as taught by Nefian, to obtain a low or minimal correlation with speech (column 2, paragraph 0016).

Allowable Subject Matter

13. **Claims 8 and 19** are objected to as being dependent upon a rejected base claim, but would be allowable if rewritten in independent form including all of the limitations of the base claim and any intervening claims.

Conclusion

14. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure.

- Brand (USPN 6,735,566) discloses generating realistic facial animation from speech.
- Basu et al. (USPN 6,219,640) discloses methods and apparatus for audio-visual speaker recognition and utterance verification.
- Nefian (USPN 7,165,029) disclose a coupled hidden markov model for audiovisual speech recognition.
- Moore (USPN 6,707,921) discloses a use of mouth position and mouth movement to filter noise from speech in a hearing aid.

15. Any inquiry concerning this communication or earlier communications from the examiner should be directed to Jakieda R. Jackson whose telephone number is 571.272.7619. The examiner can normally be reached on Monday through Friday from 7:30 a.m. to 5:00p.m.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, David Hudspeth can be reached on 571.272.7843. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

JRJ
February 8, 2007

DAVID HUDSPETH
SUPERVISORY PATENT EXAMINER
TECHNOLOGY CENTER 2600